**Final Report - Web Based Population Clustering**

By Nathan Barthen (ncb1006@sru.edu), Connor Barthen (cnb1021@sru.edu), and

Cole Cassano (cjc1029@sru.edu)

**References for Project**

**DBScan (Used with Algorithm, Cluster, DataPoint, and DBScan.java):**

 Incorrect example (but useful): <https://www.dataonfocus.com/dbscan-java-code/>

 \* Pseudocode example:

<https://www.researchgate.net/publication/325059373/figure/fig2/AS:624653831790593@1525940487951/Pseudocode-of-the-DBSCAN-algorithm.png>

**Getting School’s Lon/Lat Values**:

How to access information from url - <https://www.javatpoint.com/java-get-data-from-url>

**Reading shapefiles**:

<https://docs.geotools.org/stable/userguide/library/data/shape.html>

-Example on how to read/iterate over a .dbf file (part of a shapefile).

**CSVReader Example (library used for read csv/text files quickly**:

<https://www.journaldev.com/12014/opencsv-csvreader-csvwriter-example>

**Importing / Reading CSV Files (used for CountyDataset.java)**:

<https://www.youtube.com/watch?v=jOKsxtyZrxw&ab_channel=SlashCode>

**HTML/JavaScript**:

<https://www.computerhope.com/issues/ch000317.html>

<https://developer.mozilla.org/en-US/docs/Learn/Forms/Form_validation>

-Used with for edit/add school javascript exception handling

**SMTP Protocol Example**:

<https://www.journaldev.com/2532/javamail-example-send-mail-in-java-smtp>

-Referenced when emailing the user their routing file.

**GeoJson / HTML Display**:

<https://leafletjs.com/SlavaUkraini/>

-Example resources on how to create geoJson / pass geoJson to

OpenStreetMap/Leaflet display.

**Completion:** *Complete*

**Hindsight Changes**

-It would save on load times if we did not have to look at all the addresses in the given state when selecting addresses inside of the district’s polygon. If it were possible to find counties located near the selected county, it would be possible to only read a few county’s addresses (instead of combining all address for the state into one).

-Implement the rudimentary display sooner to ensure the clustering was working.

-Some states may not split up school districts into counties, so it may be better to remove counties, and adds a search option to select school districts.

-The NACE Dataset did not have county names so to get county names (for each school district) we used a separate database. This may lead to some school districts being left out since it may not have been in the second database (*thought almost all were added, if not all*). It may have been better to create a second program, that could be run on the NACE Dataset which would take the city name, determine which county that city is in and then add another column with the district information. This could be run and calculated, then the new excel (csv) file could be added to the database (instead of having to use two separate csv files from different sites).

**Expansion / Future Work**

-Implement add a Login / Sign Up Page. The groundwork for this is already implemented and just needs expanded upon. This would limit where users can go and what they can control. Giving every user having full unrestricted access is a security risk.

-Implement Dr. Thangiah’s Bus Routing Algorithm. This algorithm would take the cluster information and generates bus routes. Plots routes based on the most optimal paths.

-Add more states to the program to encompass the entire country. It currently only supports Pennsylvania, but the steps to add more states is detailed in the Technical Manual.

-Add additional exception handling for adding schools. More specifically, the address, city, and zip code to ensure the information is within that district’s polygon.

-Add different clustering methods when generating the clusters for the given school(s). It currently only uses DBScan clustering.

**Glossary**:

**DBScan** – Clustering algorithm which can be used to with addresses (lon/lat) to

cluster addresses into group, removing noise in the process.

**Diameter** – how big each cluster will be in size.

**Shapefile** – file type used for geospatial data (map data).

**County** – subdivision of a state (polygon) containing many school districts.

**School Districts** – subdivision of a county, home to many schools.

**Batch Processing** – running programs in batch (i.e. generating the routing files

for every school district for every county in a given state). Letting it run over a

long period of time (hours/days)

**GeoJson** – filetype / text format used for displaying polygon (with

points/addresses) to a html webpage.